UNIVERSITY OF CRAIOVA
Faculty of Automation, Computers and Electronics

Ph.D. Student Laviniu Aurelian Bădulescu

- PH.D. THESIS ABSTRACT -

„DATA MINING” DECISION TREES ALGORITHMS OPTIMIZATION

Supervisor: Acad. Prof. Univ. Dr. Ing. Mircea Petrescu

1. TABLE OF CONTENTS (Excerpts)

2. KEYWORDS: *data mining*, *classification problem, decision trees*, *training dataset, test dataset, class, class labels, splitting criteria*, *knowledge discovery, decision rules*, *unpruned decision tree, pruning methods, pruned decision tree, classification error rate*, *classification accuracy*.

3. SYNTHESIS OF THE PH.D. THESIS MAIN PARTS

Ph.D. thesis objectives targeted:

- Decision Trees (DT) algorithms optimisation, in order to find optimal splitting criteria and optimal pruning methods;
- developing a unified algorithmically framework for approaching the DT algorithms, the splitting criteria and the pruning methods;
- developing an experimental framework where, based on proposed software package for induction, pruning and execution of the DT, for computing confusion matrix and associated decision rules, the research to be done.

The Ph.D. thesis is divided into six chapters preceded by a table of contents and followed by References. The table of contents is structured in accordance with the domain's themes and it includes the following parts: Introduction (Chapter 1), State of knowledge (Chapter 2), Own contributions (Chapter 3, Chapter 4, Chapter 5), Final conclusions (Chapter 6) and References.

In **Chapter 1, Introduction**, an introduction is realized to establish the theme treated. Thus, it shows the relation between Knowledge Discovery from Databases and Data Mining (DM), the classification problem, basic concepts surrounding DT, as well as advantages and disadvantages of DT algorithms. At the end of the chapter, it presents how the thesis is organized and the original contributions of its author. One of the main personal contributions on a theoretical level is the construction of a unified frame that unitarily treats several dozen splitting criteria and, in the same mathematical framework, the presentation of a few pruning methods.

Considering that the main approaches of optimization of DT algorithms must follow two directions: choosing an optimal splitting criterion and the best pruning method, the major personal contribution of the thesis on a practical level is represented by the discovery, based on a large number of personal experiments, of a splitting criterion that systematically shows the best performances, compared to dozens of other criteria considered for experimentation. At the same time, the confidence level pruning DT method was demonstrated to be superior. Thus, new DT algorithms have been conceived, optimized, representing a personal contribution of the thesis.

Another key contribution of the PhD thesis was the development of an experimental framework which, based on a software package meant for the induction, pruning and execution of the DT, for generating the confusion matrix and the decision rules associated to DT, made numerous experiments using highly differing databases. The experimental results were compared with many results of other researchers that have experimented on the same databases, but have used different DM algorithms, finding that personal experimental results are just as accurate as the best literature values.

Another major personal contribution is the study in Chapter 2, the result of an exhaustive and up-to-date browsing of an enormous volume of works in the DT domain, in conjunction with a sustained effort of synthesizing and systematization of the reading material. This chapter shows a personal synthesis of the main DT algorithms, accomplished exclusively based to bibliography from foreign literature.

Chapter 2, The existing literature presentation about Decision Trees, makes a personal synthesis of the main DT algorithms, obtained as a result of browsing and systematization of an important bibliography of foreign literature. Thus, there are numerous DT algorithms present: the AID, THAID, CHAID and FIRM algorithms, the CART algorithm, the ID3, C4.5 and other related algorithms, lazy DT, DT algorithms from SAS package, multidimensional DT algorithms, oblique multivariate DT algorithms, omnivariate DT algorithms, oblivious DT algorithms, parallel DT algorithms, DT algorithms from WEKA package, DT with genetic algorithms, DT algorithms for data streams, orthogonal DT algorithms, hybrid DT, online adaptive DT, distributed DT algorithms as well as other DT algorithms.

Chapter 3, Decision Tree's induction and pruning algorithms and splitting criteria used in experiments, are an original contribution, proposing a unified algorithmic framework to address the DT algorithms and providing a thorough description of dozens of attribute selection methods for splitting the dataset corresponding to a node and several DT pruning methods (cost complexity pruning, reduced error pruning, minimum error pruning, pessimistical pruning, error based pruning or confidence level pruning, optimal pruning, minimum description length pruning). Chapter 3 presents proposed algorithms for induction and pruning of DT, impurity based criteria, normalized impurity based criteria, binary criteria, K2 and Bayes-Dirichlet criteria later used in experiments. At the same time, it presents the method of transformation of DT in the decision rules set and the way of treating missing values attributes.

Chapter 4, Materials and methods used in experiments, presents the 8 databases used in the experiments, whose dimensions are close to 600.000 cases, 55 attributes (continuous and nominal), 7 label classes, numerous cases with missing values and duplicate or contradictory records. Also in this chapter the proposed and utilized in experiments software package is presented: the DT induction program, the DT pruning program, the program for testing DT classification accuracy, the one for calculating the confusion matrix and the program for generating the decision rules associated to DT.

Chapter 5, Experimental results and discussions, presents the experiments performed for each of the 8 databases based on a set of identical processing. The performances of each step of processing were shown in tables and charts with comments at every step of the experiments.

1. The first step of the experiments was the DT induction based on the training dataset. In this step there were induced 28 DT, corresponding to the 28 attribute selection measures. Were taken into account the values of parameters of the building process of DT, such as: the number of nodes of the induced DT, the number of attributes necessary for the DT induction, the number of levels of the DT, the DT growth time, the size of the file containing the inducted DT. These values were compared and discussed, statistical indicators were calculated, tables and charts with comments were presented, extracting conclusions on the different behavior of the 28 DT induced with the 28 measures.

2. In the second step of the experiments, these 28 DT, in the unpruned form, were processed in order to extract from them the 28 sets of decision rules, one set for each DT obtained in step 1. Were taken into account the values of parameters of the DT decision rules extraction process, such as: the decision rules number, the time required to build the file containing the decision rules, the time required to read the file containing the DT, the size of the file containing the decision rule set. These values were compared and discussed, statistical indicators were calculated, tables and charts with comments were shown, and extracting conclusions on the different behavior of the 28 DT induced with the 28 attribute selection measures.

3. The third step of the experiments, accomplished on each database, involved the execution of the 28 DT induced in the first step. These experiments verified the classification model represented by the DT based on a test dataset, unknown dataset in step 1, when DT was generated. It is the most important step of the experiments, because now the values of the

classification error rate on test data or model accuracy are obtained. For each database 28 values of the classification accuracy of the unpruned DT on test data are obtained, corresponding to each of the 28 measures with which the DT was induced. By comparing these values we can draw conclusions about the performance of every attribute selection measures.

4. The fourth step of the experiments targeted the pruning of the 28 DT obtained in step 1, with the confidence level pruning method. Were taken into account the values of some parameters of the DT pruning process: the confidence level, the number of attributes necessary to build the pruned DT, the number of nodes and the number of levels of the pruned DT, the pruning time of the file containing the unpruned DT, the size of the file containing the pruned DT. These values were compared and discussed, statistical indicators were calculated, tables and charts with comments were shown, and extracting conclusions on the different behavior of the 28 confidence level pruned DT.

5. The fifth step of the experiments involved the extraction of the decision rules from the 28 DT reduced with the confidence level pruning method in step 4. The processing resembles those in step 2.

6. The sixth step involved the execution on test data of the 28 confidence level pruned DT obtained on step 4. The values of the classification error rate on the test data were retained, values which were later compared with values obtained in the execution of the unpruned DT on test data (see step 3) and with values obtained in the execution of the pessimistic pruned DT on test data (see step 7).

7. The seventh step of the experiments involved the pruning of the 28 DT, obtained in step 1, with the pessimistic pruning method. The processing resembles those in step 4.

8. The eighth step of the experiments involved the extraction of the decision rules from the 28 DT pessimistically pruned in step 7. The processing resembles those in step 2.

9. The ninth step involved the execution on the test data of the 28 DT pessimistically pruned in step 7. The processing resembles those in steps 3 and 6.

10. In the tenth step a discussion was realized on the values of the number of decision rules and the classification error rate values on test data for the three types of DT: unpruned, pruned with confidence level method and pruned with pessimistic pruning method, showing which attribute selection measure and which pruning method had the best behavior in the experiments.

11. Finally, in the eleventh step, the results of some representative works from literature were reviewed, presenting performance tests on each database and, in the same time, we made a comparison of the results obtained in those experiments with the results obtained by the 28 criteria and the two pruning methods. For reasons of space, the confusion matrix was calculated for only one database (*Image*).

In **Chapter 6, Conclusions of the experiments and future directions** are presented. Thus, three techniques are used to compare the classification performances of the 28 induced and pruned DT with two pruning methods on the 8 databases.

The first time the comparison is realized based on the values of the arithmetic mean and based on the values of the standard deviation of the classification error rate on test data for all types of DT, all databases and all attribute selection measures. Finds that the best classify DT induced with the *rciq* (quadratic information gain ratio) measure, with an average of classification error rate on test data of 18.59% (also having the smallest value for standard deviation: 14.11), followed by the DT induced with *rel* (relief), *k2*, *bd* (Bayes-Dirichlet), *lmdfa* (minimum description length in absolute frequencies) and *csh* (stochastic complexity) measure with small values for standard deviation, while the smallest average performance is achieved by DT induced with *ciqp* (balanced quadratic information gain) measure, with an average value of 31.12%.

DT induced by *rciq* measure has the best value of the mean classification error rate and the smallest standard deviation, resulting in the smallest spread of classification error rate values on test data around their mean. This result shows that the performance of DT induced by *rciq* measure is not affected too much by the features of the database, i.e. one of the fundamental characteristics of the *rciq* measure is its independence from the domain, a very important feature today, when databases are made up of data, with attributes pertaining to different domains, collected together. The diversity of domains from the composition of the databases is one of the reasons that have increased the need to use tools for automated knowledge discovery from databases and inductive learning algorithms.

The results of the experiments show that the smallest mean value (1,147.54) of the number of decision rules is achieved by the DT induced by *mapd* measure, and the worst mean performance value (1,641.25) is achieved by the DT induced by *rcs* measure. To note that the DT induced by *rciq* measure, which has the best behavior regarding classification accuracy on test data, is placed on fourth spot in the hierarchy of the smallest mean values for the number of decision rules.

The second comparison method of the classification performance values of the 28 DT was represented by the *win/tie/loss* technique. Also in this case, the performances of the DT induced by *rciq* measure are always the best. The DT induced by *rciq* measure presents the largest number of cases with a better performance value than all the other DT induced by the other measures and at the same time, the smallest number of cases with a lower performance value than the other DT. The next five places are occupied by induced DT with another 5 measures *csh*, *bd*, *k2*, *lmdfa* and *rel*, which no longer present the property shown by the *rciq* measure, of having both a large number of cases when they overcome the performance values of induced DT with all other measures, and a small number of cases when they are overcome by the performance values of the other DT.

The third comparison method of the classification performance values of the 28 DT was represented by the geometric mean of the classification error rate ratio. The performances of the DT induced by *rciq* measure surpass, in all the cases considered in the experiments, the performances of the DT induced with the other 27 measures. Is followed, in order, by the performance values achieved by the DT induced with the *rel*, *gim* (modified gini index), *bd*, *k2* *lmdfa* and *csh* measures. Thus, we can conclude that, regardless of the comparison criterion considered, optimizing the DT algorithms by using the *rciq* measure induces the most efficient DT. Even the optimization based on the *rel*, *k2*, *bd*, *lmdfa* and *csh* measures produces DT with good performances.

Regarding the search for the best pruning method, analysis of the values obtained from previous experiments presented in the thesis shows that the DT pruned with the confidence level pruning method systematically obtains the best values for classification accuracy on test data, simultaneously generating a small number of decision rules. In second place are the performances of the pessimistically pruned DT and in last place the unpruned DT performances. Of note is this differentiation, which always places confidence level pruned DT in first place and the pessimistically pruned DT in second place, it is much clearer in regard to the number of decision rules than the classification accuracy on test data. At the same time, of further note is that the pruning improves the DT performances.

The experiments that we wish to carry out in the near future will involve much larger databases, with many attributes and cases, which we wish to verify the conclusions obtained for the databases utilized until now. At the same time, we will include in the set of measures on the basis of which the splitting of a node is realized other the attribute selection criteria and we will utilize other DT pruning methods, outside of the two used in our experiments. At the same time, we will have to take into account the information provided by the confusion matrix.